Proyecto Final

DSII: Machine learning para la ciencia de datos.

CODERHOUSE

Alumno: Jorge Rodriguez

Profesor: Ignacio Russo Locatti

Tutor: Emanuel Tevez

Comisión: 61160

Adidas es una de las marcas más icónicas y reconocidas a nivel mundial, conocida principalmente por su relación con el deporte, pero también por su influencia en la moda, la cultura urbana y la tecnología aplicada al calzado y la ropa deportiva.

Desde sus primeros días, Adidas ha sido una marca profundamente vinculada con el mundo del deporte. La empresa ha patrocinado a algunos de los más grandes atletas y equipos del mundo, incluyendo a figuras como Lionel Messi, Karim Benzema, y la selección Argentina de fútbol, entre otros. También ha estado presente en algunos de los momentos más importantes de la historia del deporte, como en las primeras Copas del Mundo de fútbol y los Juegos Olímpicos.

Adidas es mucho más que una marca deportiva. Es un emblema de innovación, estilo y compromiso con el futuro. A lo largo de su historia, ha sabido mantenerse a la vanguardia tanto en el ámbito tecnológico como en la moda, siempre con el deporte como núcleo. Hoy en día, es una de las marcas más importantes del mundo, no solo en el ámbito deportivo, sino también en la moda y la cultura global.

En 2020 vivimos un acontecimiento atípico, la pandemia mundial nos obligo no solo a quedarnos en casa sino que cambio nuestra forma de vida. Tuvimos que mantener distancia, precauciones y hasta cambiar nuestra forma de trabajar o ir de compras.

En esto último se basa este dataset, que está compuesto por las ventas de Adidas en Estados Unidos en los año 2020 y 2021.

Este trabajo buscara demostrar el impacto de la pandemia en las ventas, cantidades, métodos y cuáles son las categorías más vendidas. El impacto de los días festivos e intentaremos predecir mediante técnicas de Machine Learning las ventas futuras.

EDA Y DATA WRANGLING

Si bien es un dataset muy completo hubo que hacer algunas modificaciones.

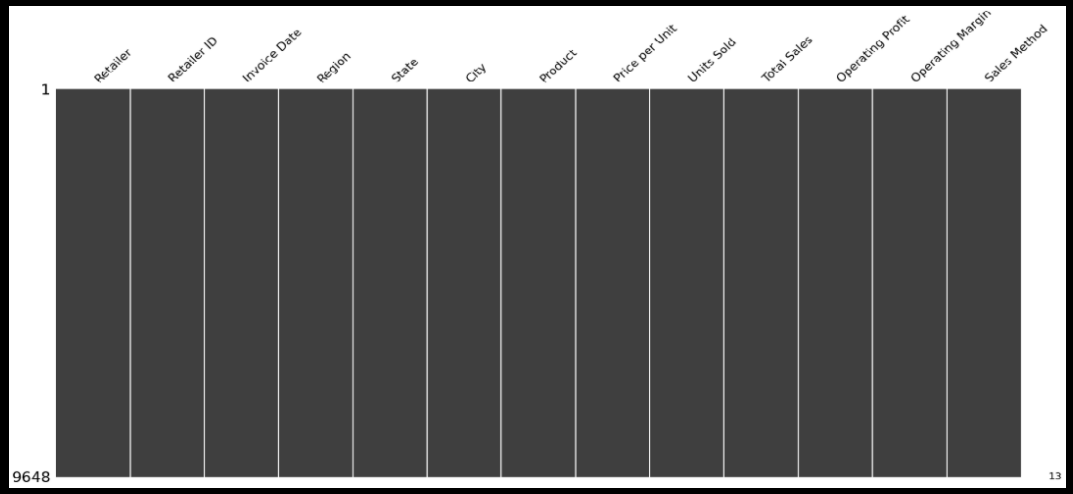
En principio las primeras filas estaban libres o no tenían datos relevantes así que se eliminaron, se utilizo la siguiente como índice y también se elimino la primera columna que también estaba en blanco.

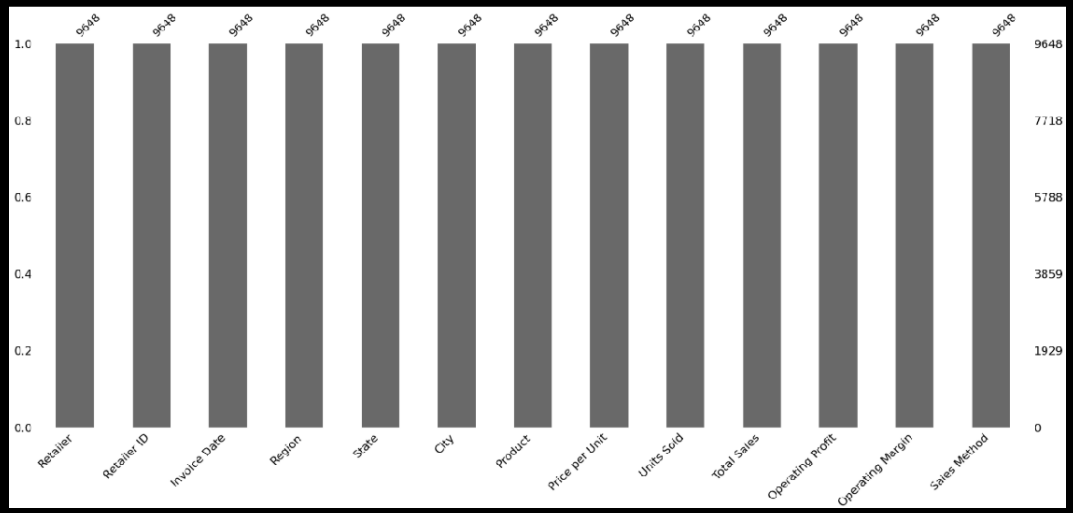
```
url = url.iloc[3:]
url = url.rename(columns=url.iloc[0]).drop(url.index[0])
url = url.drop(url.columns[0], axis=1)
```

Mediante la función .info() pudimos ver las columnas, valores nulos y el tipo de datos.

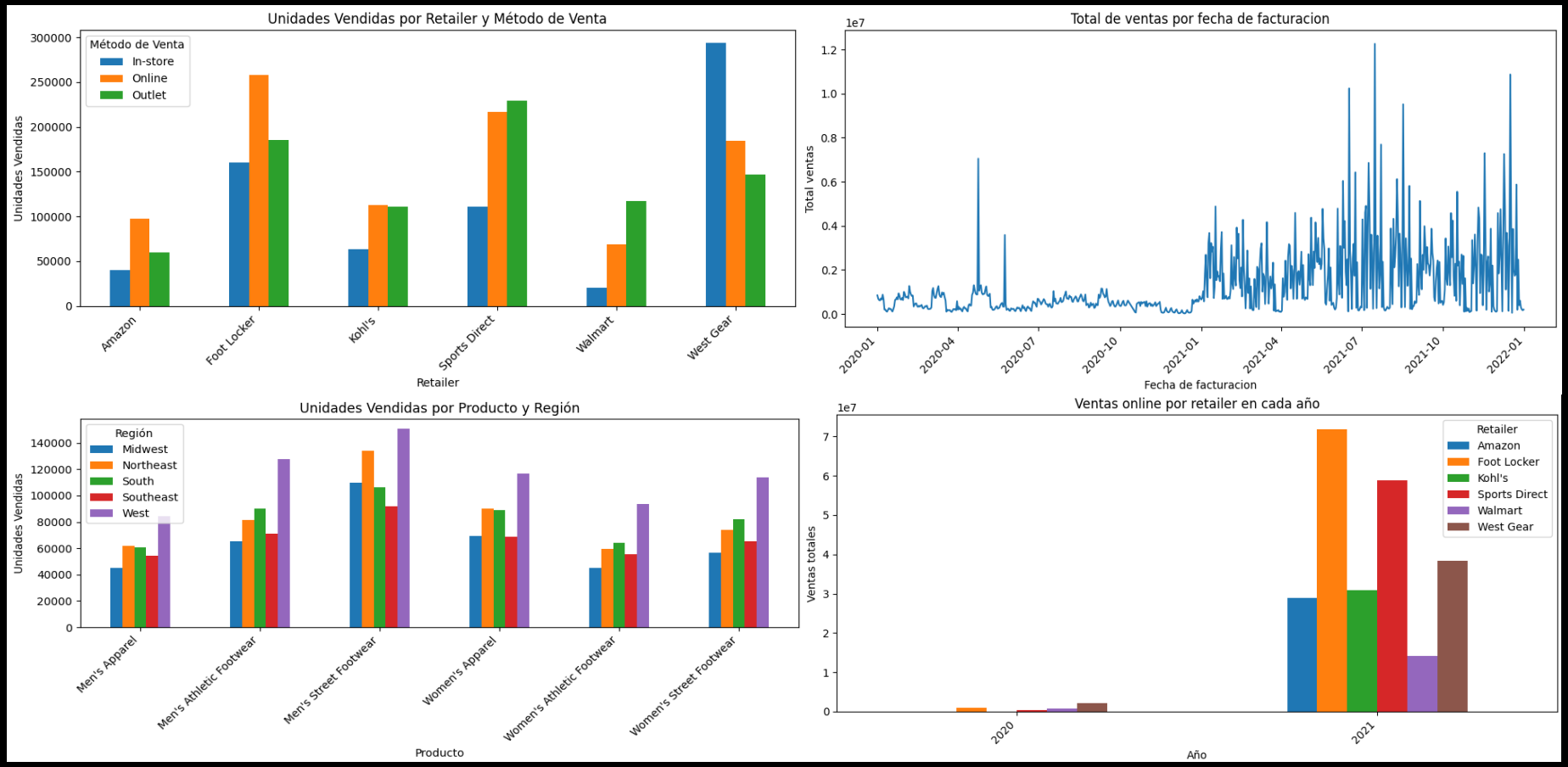
```
RangeIndex: 9648 entries, 4 to 9651
Data columns (total 13 columns):
 #   Column            Non-Null Count  Dtype 
---  ------            --------------  ----- 
 0   Retailer          9648 non-null   object
 1   Retailer ID       9648 non-null   object
 2   Invoice Date      9648 non-null   object
 3   Region            9648 non-null   object
 4   State             9648 non-null   object
 5   City              9648 non-null   object
 6   Product           9648 non-null   object
 7   Price per Unit    9648 non-null   object
 8   Units Sold        9648 non-null   object
 9   Total Sales       9648 non-null   object
 10  Operating Profit  9648 non-null   object
 11  Operating Margin  9648 non-null   object
 12  Sales Method      9648 non-null   object
```

Claramente se veía que no había nulos pero igual instalamos y corrimos la librería missingno.



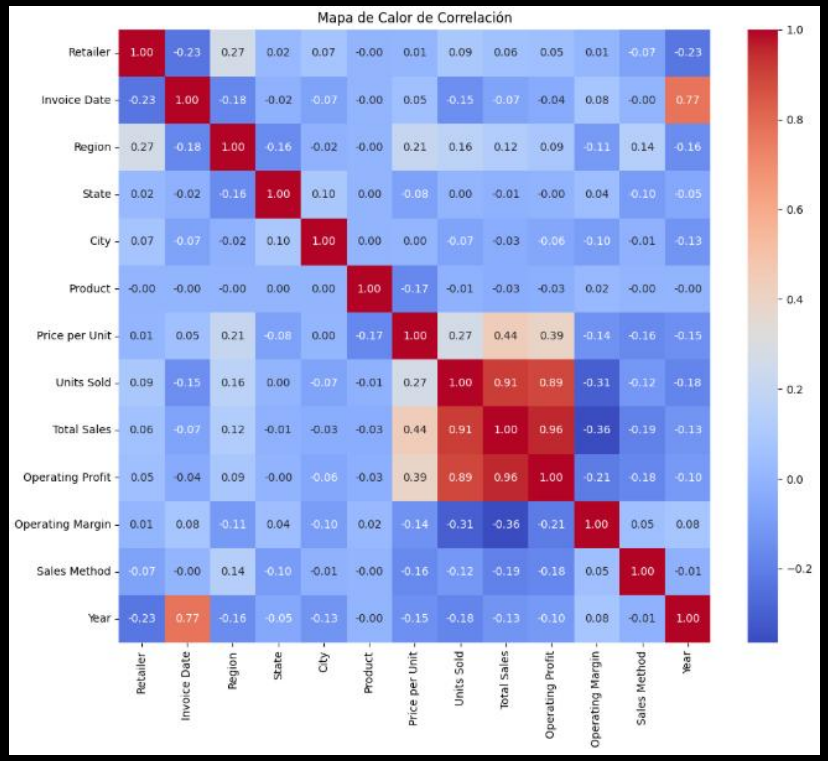


Visualizaciones



En estos gráficos podemos ver cuánto vendió cada retailer y en que modalidad, las ventas por fecha, las unidades vendidas de cada producto por región y las unidades vendidas por retailer por año.

Para ver la correlación entre variables realice un mapa de calor, el cual arrojo que las que más se relacionan entre si son ‘Units sold’, ‘Total sales’ y ‘Operating profit’.



Esto nos demuestra que a mayores unidades vendidas, mayores son las ventas totales y mayor es la ganancia.

Algoritmos de Machine Learning

En primer lugar hice un algoritmo de regresión lineal para predecir las ventas en el año siguiente, solo tome en cuenta el año 2021ya que el 2020 fue un año atípico y los resultados no serian acertados.

A continuación realice un árbol de decisión, también con los datos de 2021.

Para tener un informe mas completo
entrene también los modelos de
‘Random Forest’, ‘K-Nearest Neighbors’ y
‘SVM’ con sus respectivas validaciones.

Los métodos de validación fueron MSE y R^2

```
Modelo: Random Forest
MSE (2021): 255.9047076047905
R^2 (2021): 0.9997812410462047
--------------------
Modelo: K-Nearest Neighbors
MSE (2021): 210738.1180598803
R^2 (2021): 0.8198514960390706
--------------------
Modelo: Linear Regression
MSE (2021): 378096.37808194605
R^2 (2021): 0.67678606276082226
--------------------
Modelo: SVM
MSE (2021): 1170081.3724495524
R^2 (2021): -0.0002386407353380804
```

Teniendo en cuenta estos métodos de validación el modelo que mejor se ajusta es Random Forest, seguido de KNN.

Por la alta puntuación que devolvió el coeficiente de determinación puede interpretarse que presente Overfitting pero considero no es el caso porque la intención era predecir las ventas del año siguiente solo con un año de registros. Era esperable que de así.

Hypertunning de Parámetros

Como punto necesario para esta entrega se solicito hacer este paso. En mi caso el modelo dio muy bien y no era necesario, incluso el resultado que arrojo no fue mejor que el original. Por esto y para no perder tiempo con algo que no iba a necesitar, elegí hacer un Random grid.

Conclusión

Mediante las visualizaciones se pudo demostrar las cantidades vendidas por retailer y en que modalidad, cuales son los productos más vendidos y las regiones que más venden, las fechas en las que aumentan las ventas y la gran diferencia que hubo entre 2020 y 2021en las ventas totales.

En respuesta a la pregunta inicial, si se pueden predecir las ventas para el año siguiente fue determinante probar distintos modelos para encontrar uno que devuelva un resultado satisfactorio pero se pudo realizar.